Global Convergence of Least Squares EM for Demixing Two Log-Concave Densities

Wei Qian, Yuqian Zhang, Yudong Chen {wq34,yz2557,yudong.chen}@cornell.edu

Cornell Engineering
Operations Research and Information Engineering

Motivation

The **Expectation Maximization** (EM) algorithm has only a few theoretical guarantees for convergence despite its popularity.

- Recent progress on global convergence has focused on a balanced mixture of 2 Gaussian distributions.
- Can we develop a global convergence theory for a mixture of some broader class of distributions?

Problem Overview

- **Distribution Class**: log-concave and rotation invariant

$$\mathcal{F} = \Big\{ f : f(\boldsymbol{x}) = \frac{1}{C_g}\exp\big(-g(\|\boldsymbol{x}\|_2)\big),\ g \text{ convex and increasing on } [0,\infty),\ \int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1,\ \int x_i^2 f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1, \forall i \in [d] \Big\}.$$

Each $f \in \mathcal{F}$ generates a location-scale family consisting of the densities

$$f_{\beta,\sigma}(\boldsymbol{x}) := \frac{1}{\sigma^d} f\left(\frac{\boldsymbol{x}-\beta}{\sigma}\right)$$

- **A Balanced 2-Mixture Generative Model**

$$D(\beta^*,\sigma) := \frac{1}{2} f_{\beta^*,\sigma} + \frac{1}{2} f_{-\beta^*,\sigma}.$$

- **Location Estimation Problem**: Given data $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n \in \mathbb{R}^d$ sampled i.i.d. from the mixture distribution $D(\beta^*,\sigma)$, σ is known, how to estimate β^*?
- **Classical EM** does not have a closed-form solution for the M-step.
- We analyze the **Least-Squares EM** (LS-EM):
  - **E-step:** Compute the conditional probabilities given β,

$$p^1_{\beta,\sigma}(\boldsymbol{X}) := \frac{f_{\beta,\sigma}(\boldsymbol{X})}{f_{\beta,\sigma}(\boldsymbol{X}) + f_{-\beta,\sigma}(\boldsymbol{X})},$$

$$p^2_{\beta,\sigma}(\boldsymbol{X}) := \frac{f_{-\beta,\sigma}(\boldsymbol{X})}{f_{\beta,\sigma}(\boldsymbol{X}) + f_{-\beta,\sigma}(\boldsymbol{X})}.$$

  - Least-Squares M-step: weighted least squares regression

$$M(\beta^*,\beta) = \operatorname*{argmin}_{\boldsymbol{b}} \mathbb{E}_{\boldsymbol{X}\sim D(\beta^*,\sigma)} \left[p^1_{\beta,\sigma}(\boldsymbol{X}) \|\boldsymbol{X}-\boldsymbol{b}\|_2^2 + p^2_{\beta,\sigma}(\boldsymbol{X}) \|\boldsymbol{X}+\boldsymbol{b}\|_2^2 \right]$$

$$= \mathbb{E}_{\boldsymbol{X}\sim D(\beta^*,\sigma)} \boldsymbol{X} \tanh\left(\frac{1}{2} g\left(\frac{1}{\sigma}\|\boldsymbol{X}+\beta\|_2\right) - \frac{1}{2} g\left(\frac{1}{\sigma}\|\boldsymbol{X}-\beta\|_2\right) \right).$$

Properties of Least Squares EM iterates

- **Two Dimensional Structure**: The LS-EM iterate $M(\beta^*,\beta)$ is in the span of β and β^*.
  - Invariant 1-dim subspace: in the direction of β^* or in the orthogonal direction to β^*.
- **Angle Decreasing Property**: The angle between the LS-EM iterate $M(\beta^*,\beta)$ and $\operatorname{sign}(\langle\beta,\beta^*\rangle)\beta^*$ is smaller than the angle between β and $\operatorname{sign}(\langle\beta,\beta^*\rangle)\beta^*$ when β is not orthogonal to β^*.

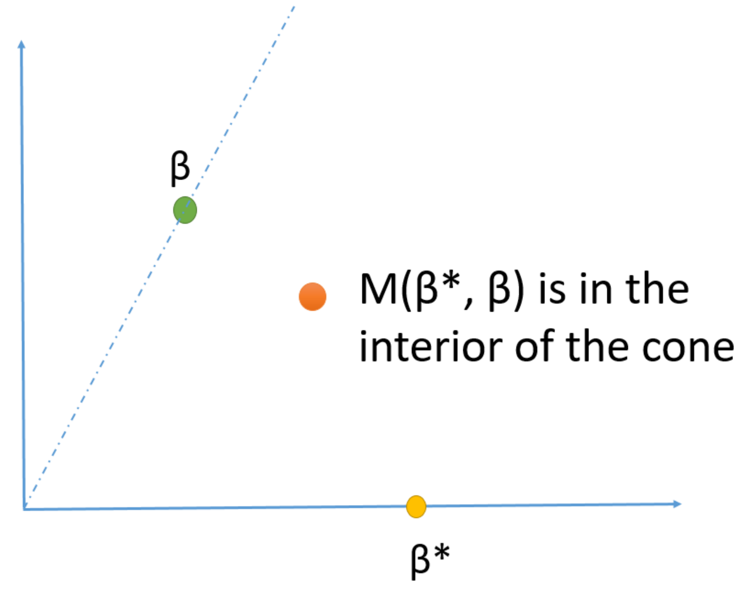

Figure: The LS-EM iterate has a smaller angle with β^*.

- **Asymptotic Convergence**: The Least-Squares EM algorithm converges to $\operatorname{sign}(\langle\beta^0,\beta^*\rangle)\beta^*$ from any randomly initialized point β^0 that is not orthogonal to β^*
  - The angle decreasing property forces the iterates to converge to the correct subspace;
  - The dynamics along the β direction forces the iterates to converge to the ground truth.
- **Explicit convergence rate** in 1-D case, $z = \min(|\beta|, |\beta^*|)$:

$$|M(\beta^*,\beta) - \operatorname{sign}(\beta\beta^*)\beta^*| \le \kappa(\beta^*,\beta,\sigma) \cdot |\beta - \operatorname{sign}(\beta\beta^*)\beta^*| \tag{1}$$

  - Gaussian: $\kappa(\beta^*,\beta,\sigma) \le \exp\big(-z^2/2\sigma^2\big)$,
  - Laplace: $\kappa(\beta^*,\beta,\sigma) \le \frac{2\exp(-\frac{\sqrt{2}}{\sigma}z)}{1+\exp(-2\frac{\sqrt{2}}{\sigma}z)}$
  - Logistic: $\kappa(\beta^*,\beta,\sigma) \le \frac{4\exp(-\frac{\pi z}{\sigma\sqrt{3}})}{1+\exp(-\frac{2\pi z}{\sigma\sqrt{3}})+2\exp(-\frac{\pi z}{\sigma\sqrt{3}})}$

Different Behaviors Compared to 2GMM

- In 1-D, the contraction in distance to the ground truth (1) holds for all $f \in \mathcal{F}$;
- In higher dimension, the contraction in ℓ_2 distance still holds for Gaussian. However, there exists some log-concave distribution such that the ℓ_2 distance strictly increases.

Robustness under Model Mis-specification

In practice, we do not know f that generates the data. Instead, we fit with some $\widehat{f} \in \mathcal{F}$. The LS-EM iterate under the mis-specification setting is:

$$\widehat{M(\beta^*,\beta)} = \mathbb{E}_{\boldsymbol{X}\sim D(\beta^*,\sigma)} \boldsymbol{X} \tanh\left(\frac{1}{2}\widehat{g}\left(\frac{1}{\sigma}\|\boldsymbol{X}+\beta\|_2\right) - \frac{1}{2}\widehat{g}\left(\frac{1}{\sigma}\|\boldsymbol{X}-\beta\|_2\right) \right).$$

- **Preserved properties**: two dimensional structure; angle decreasing property.
- **3-fixed points**: there are only 3 fixed points $\{\pm\overline{\beta}, 0\}$ in the direction of β^*.
- **Using Gaussian is a good choice:** when $\widehat{f}$ is Gaussian, $|\overline{\beta} - \beta^*| <= 10\sigma$ if the SNR $|\beta^*|/\sigma$ is moderate.

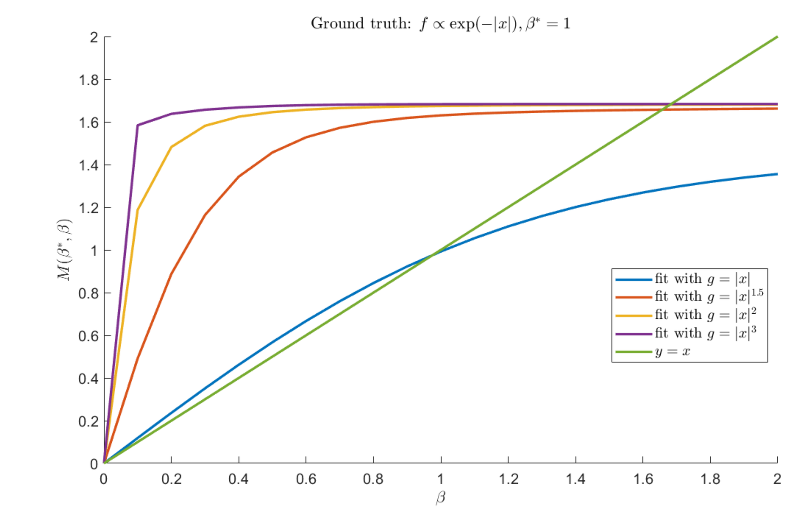

Figure: Ground truth: $f \propto \exp(-|x|)$, plot of $\widehat{M}(\beta^*,\beta)$ with $g = |x|, |x|^{1.5}, |x|^2$ and $|x|^3$.

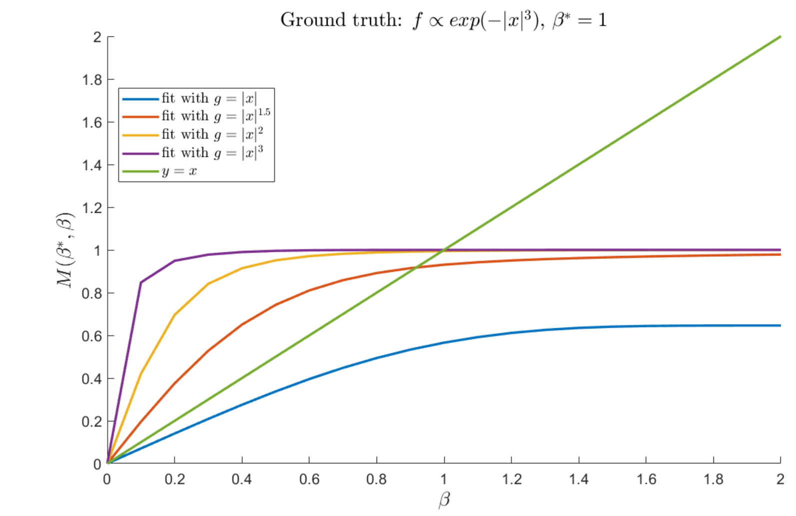

Figure: Ground truth: $f \propto \exp(-|x|^3)$, plot of $\widehat{M}(\beta^*,\beta)$ with $g = |x|, |x|^{1.5}, |x|^2$ and $|x|^3$.

Summary

- Two dimensional structure of the Least Squares EM under both correctly specified and mis-specified settings. Rotation invariance assumption guarantees this property;
- Angle decreasing of the Least Squares EM: convergence to the correct subspace. Log-concavity assumption and monotonicity of g guarantee this property.